SOaN : un algorithme pour la coordination d'agents apprenants et non communicants.

Laëtitia Matignon, Guillaume J. Laurent, Nadine Le Fort-Piat

SOaN: un algorithme pour la coordination d'agents apprenants et non communicants

Laëtitia Matignon, Guillaume J. Laurent, Nadine Le Fort-Piat

Institut FEMTO-ST, Département AS2M
UFC-ENSMM-UTBM-CNRS, Université de Franche-Comté, Besançon, France
laetitia.matignon@ens2m.fr, guillaume.laurent@ens2m.fr,
nadine.piat@ens2m.fr
http://www.femto-st.fr/

Mots-clés : Apprentissage par renforcement, systèmes multi-agents, jeux de Markov d'équipe, agents non communicants

L'apprentissage par renforcement dans les systèmes multi-agents est un domaine de recherche très actif, comme en témoignent les états de l'art récents [Busoniu *et al.*, 2008, Sandholm, 2007, Bab & Brafman, 2008, Vlassis, 2007]. Lauer et Riedmiller ont notamment montré que, sous certaines hypothèses, il est possible à des agents apprenants simultanément de coordonner leurs actions sans aucune communication et sans qu'ils perçoivent les actions de leurs congénères [Lauer & Riedmiller, 2000]. Cette propriété est particulièrement intéressante pour trouver des stratégies de coopération dans les systèmes multi-agents de grande taille.

Néanmoins, l'apprentissage d'agents non communicants soulève plusieurs difficultés qu'aucun algorithme ne surmonte complètement. La première difficulté concerne la remise en cause de la propriété fondamentale de Markov au niveau local [Bowling & Veloso, 2000]. La seconde est liée à la présence éventuelle dans le processus de facteurs rendant la coordination difficile (bruits, équilibres multiples, équilibres cachés) [Panait *et al.*, 2008] . Enfin, l'exploration des agents a un impact souvent délétère sur l'apprentissage dans le cas d'agents non communicants [Matignon *et al.*, 2009]. Trouver un algorithme robuste à la stratégie d'exploration est un enjeu important qui est peu évoqué dans la littérature et rarement pris en compte dans les algorithmes d'apprentissage par renforcement décentralisés.

Parmi les approches proposées dans la littérature pour l'apprentissage d'agents non communiquants, le Q-learning distribué [Lauer & Riedmiller, 2000] est le seul algorithme qui possède une preuve de convergence vers un équilibre de Nash Pareto optimal dans les jeux de Markov d'équipe [1] déterministes. Il est fondé sur la notion d'agents optimistes qui sont robustes à l'exploration des autres agents. Son unique handicap est de ne s'appliquer qu'aux environnements déterministes. D'autres algorithmes ont été proposés pour les jeux de Markov d'équipes stochastiques. On peut notamment citer le Q-learning indicé [Lauer & Riedmiller, 2004], le Wolf-PHC [Bowling & Veloso, 2002], le Q-Learning hystérétique [Matignon *et al.*, 2007] et le FMQ récursif [Matignon *et al.*, 2008]. Le Q-learning indicé comporte des garanties théoriques mais n'est pas applicable en pratique à cause de l'explosion combinatoire de la taille des listes utilisées vis-à-vis du nombre de couples état-action. Le Wolf-PHC quant à lui ne surmonte pas les équilibres cachés. Le Q-Learning hystérétique est peu robuste à la stratégie d'exploration. Enfin, le FMQ récursif ne s'applique qu'aux jeux matriciels [2].

Nous proposons un nouvel algorithme, le *Swing between Optimistic and Neutral* (SOaN) qui a pour objectif de répondre à tous ces enjeux [Matignon *et al.*, 2009]. Cet algorithme est fondé sur une estimation récursive de la probabilité de réalisation de la politique optimiste des agents. Cette estimation permet d'évaluer

1. Un jeu de Markov d'équipe est un jeu de Markov dans lequel tous les joueurs reçoivent la même récompense. En outre, par définition, l'observabilité de l'état y est totale.

2. Jeu de Markov a un seul état.

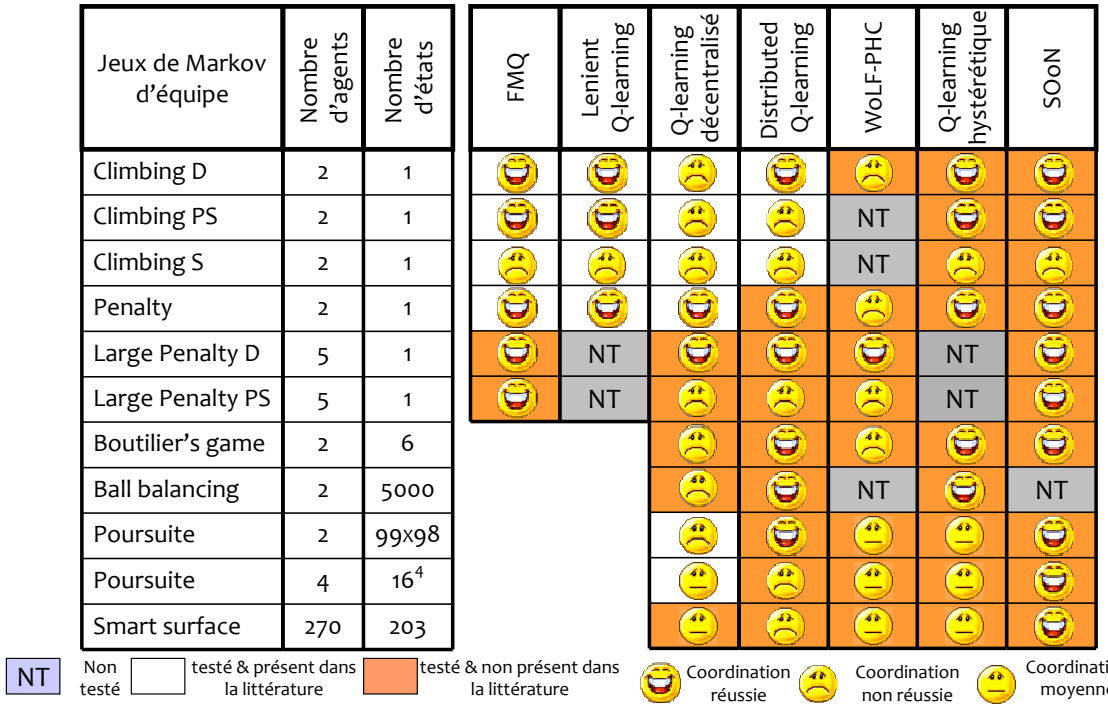

TABLE 1 – Résultats des tests du SOaN et d'algorithmes de la littérature sur différents benchmarks.

les couples état-action à l'aide d'une interpolation linéaire entre les évaluations moyennes du Q-learning [Watkins & Dayan, 1992] et les évaluations optimistes du Q-learning distribué [Lauer & Riedmiller, 2000]. Cette heuristique permet aux évaluations des actions de se positionner automatiquement entre les évaluations optimistes et moyennes selon la stochasticité détectée dans le processus. Au départ, le processus est supposé déterministe et les agents sont optimistes. Puis, l'algorithme s'adapte automatiquement à la « stochasticité » de son environnement à travers l'estimation de la probabilité de réalisation de la politique optimiste. L'algorithme SOaN surmonte ainsi les principaux facteurs de non-coordination évoqués plus haut et est de plus robuste face à l'exploration des autres agents.

Les simulations effectuées sur une douzaine de benchmarks multi-agents montrent la présence d'une première phase d'adaptation automatique des évaluations avant une phase finale de coordination. Elles confirment ainsi que l'interpolation linéaire est une bonne heuristique d'évaluation des actions. Comparé aux autres algorithmes de la littérature sur les mêmes benchmarks [3] (cf. table 1), le SOaN se comporte mieux ou aussi bien, tout en présentant une meilleure robustesse et une plus grande facilité de réglage.

Références

BAB A. & BRAFMAN R. I. (2008). Multi-agent reinforcement learning in common interest and fixed sum stochastic games : An experimental study. *Journal of Machine Learning Research*, p. 2635–2675.

BOWLING M. & VELOSO M. (2002). Multiagent learning using a variable learning rate. *Artificial Intelligence*, **136**, 215–250.

BOWLING M. & VELOSO M. M. (2000). *An analysis of stochastic game theory for multiagent reinforcement learning*. Rapport interne CMU-CS-00-165, Computer Science Department, Carnegie Mellon University.

BUSONIU L., BABUSKA R. & SCHUTTER B. D. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C : Applications and Reviews*, **3838**(2), 156–172.

3. Tous ces essais ont été réalisés avec la bibliothèque libre BOSAR (`www.lab.cnrs.fr/openblockslib/`).

LAUER M. & RIEDMILLER M. (2000). An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proc. of the Int. Conf. on Machine Learning*, p. 535–542 : Morgan Kaufmann.

LAUER M. & RIEDMILLER M. (2004). Reinforcement learning for stochastic cooperative multi-agent systems. In *Proc. of the Int. Conf. on Autonomous Agents and Multiagent Systems*, p. 1516–1517.

MATIGNON L., LAURENT G. J. & FORT-PIAT N. L. (2008). A study of fmq heuristic in cooperative multi-agent games. In *Proc. of the Int. Conf. on Autonomous Agents and Multiagent Systems, Workshop 10 : Multi-Agent Sequential Decision Making in Uncertain Multi-Agent Domains*.

MATIGNON L., LAURENT G. J. & LE FORT-PIAT N. (2007). Hysteretic q-learning : an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *Proc. of the IEEE Int. Conf. on Intelligent Robots and Systems*, p. 64–69, San Diego, CA, USA.

MATIGNON L., LAURENT G. J. & LE FORT-PIAT N. (2009). *Coordination of independent learners in cooperative Markov games*. Rapport interne RR-2009-01, Institut FEMTO-ST/UFC-ENSMM-UTBM-CNRS, Besançon, France. http ://hal.archives-ouvertes.fr/docs/00/37/08/89/PDF/Rapport-1.pdf.

PANAIT L., TUYLS K. & LUKE S. (2008). Theoretical advantages of lenient learners : An evolutionary game theoretic perspective. *Journal of Machine Learning Research*, **9**, 423–457.

SANDHOLM T. (2007). Perspectives on multiagent learning. *Artificial Intelligence*, **171**, 382–392.

VLASSIS N. (2007). A concise introduction to multiagent systems and distributed artificial intelligence. In R. BRACHMAN & T. DIETTERICH, Eds., *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan & Claypool.

WATKINS C. J. & DAYAN P. (1992). Technical note : Q–learning. *Machine Learning*, **8**, 279–292.